PROCEEDINGS B

rspb.royalsocietypublishing.org

Research

Cite this article: Bråte J, Adamski M, Neumann RS, Shalchian-Tabrizi K, Adamska M. 2015 Regulatory RNA at the root of animals: dynamic expression of developmental lincRNAs in the calcisponge *Sycon ciliatum*. *Proc. R. Soc. B* **282**: 20151746.
http://dx.doi.org/10.1098/rspb.2015.1746

Received: 21 July 2015
Accepted: 18 November 2015

Subject Areas:
evolution, genetics, developmental biology

Keywords:
lncRNA, lincRNA, regulatory RNA, sponges, development, evolution

Authors for correspondence:
Jon Bråte
e-mail: jon.brate@ibv.uio.no
Maja Adamska
e-mail: maja.adamska@anu.edu.au

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2015.1746 or via http://rspb.royalsocietypublishing.org.

THE ROYAL SOCIETY PUBLISHING

Regulatory RNA at the root of animals: dynamic expression of developmental lincRNAs in the calcisponge *Sycon ciliatum*

Jon Bråte[1], Marcin Adamski[2,3], Ralf S. Neumann[1], Kamran Shalchian-Tabrizi[1] and Maja Adamska[2,3]

[1]Centre for Epigenetics, Development and Evolution (CEDE), Department of Biosciences, University of Oslo, Oslo, Norway
[2]Sars International Centre for Marine Molecular Biology, University of Bergen, Bergen, Norway
[3]Research School of Biology, Australian National University, Canberra, Australian Capital Territory, Australia

JB, 0000-0003-0490-1175

Long non-coding RNAs (lncRNAs) play important regulatory roles during animal development, and it has been hypothesized that an RNA-based gene regulation was important for the evolution of developmental complexity in animals. However, most studies of lncRNA gene regulation have been performed using model animal species, and very little is known about this type of gene regulation in non-bilaterians. We have therefore analysed RNA-Seq data derived from a comprehensive set of embryogenesis stages in the calcareous sponge *Sycon ciliatum* and identified hundreds of developmentally expressed intergenic lncRNAs (lincRNAs) in this species. *In situ* hybridization of selected lincRNAs revealed dynamic spatial and temporal expression during embryonic development. More than 600 lincRNAs constitute integral parts of differentially expressed gene modules, which also contain known developmental regulatory genes, e.g. transcription factors and signalling molecules. This study provides insights into the non-coding gene repertoire of one of the earliest evolved animal lineages, and suggests that RNA-based gene regulation was probably present in the last common ancestor of animals.

1. Introduction

Long non-coding RNAs (lncRNAs) are usually defined as RNA transcripts which are several hundred nucleotides long but have no obvious protein-coding potential, although, in some cases, they might be translated, yielding short peptides of unknown function [1,2]. lncRNAs can regulate the expression of other genes through a variety of different mechanisms. The gene regulatory power of lncRNAs lies in their ability to interact with DNA in a site-specific manner, and at the same time bind to different proteins, bridging chromosomes and protein complexes [1,3–5]. Most nuclear lncRNAs function by guiding chromatin modifying proteins to specific genomic positions and can sometimes organize entire chromosomes or epigenetically alter chromosome states [6–8]. On the other hand, cytoplasmic lncRNAs regulate translation and stability of coding transcripts as well as protein localization (reviewed in [9]).

The lncRNA category comprises a wide variety of RNA transcripts, including both polyadenylated and non-polyadenylated lncRNAs that may be sense or antisense, intronic and intergenic with respect to protein-coding genes [10]. However, most studies on lncRNAs focus on polyadenylated lncRNAs that do not overlap other protein-coding genes, the so-called long 'intervening', or 'intergenic', non-coding RNAs (lincRNAs; [11]). lincRNAs seem to be expressed in a more tissue-specific and developmental stage-specific manner than protein-coding genes; in fact, embryonic development seems to be a very active time for lincRNA expression [12–15].

The action of lincRNAs during development has mostly been investigated in model vertebrate species. In zebrafish, a large number of lincRNAs are expressed during embryogenesis [16], and developmental regulatory functions have been demonstrated for two lincRNAs tested in knock-down and rescue by overexpression experiments [11]. In mice, more than a thousand lincRNAs are differentially expressed during postnatal testis development [17], and many lincRNAs are essential for survival and correct brain development [18]. Developmental lncRNAs have also been identified among invertebrates, for example in the nematode *Caenorhabditis elegans* [19], and recently Gaiti *et al.* [15] described dynamically expressed lncRNAs across multiple developmental stages of the demosponge *Amphimedon queenslandica*.

It has been hypothesized that an RNA-based gene regulation was important for the evolution of increased developmental complexity in animals [1]. However, it is currently not known whether this mode of gene regulation is exclusive to bilaterian animals, or whether this 'hidden layer' of gene regulation was already present in the earliest evolved (i.e. non-bilaterian) animal lineages. The findings of Gaiti *et al.* [15] based on embryonic and postembryonic development of *A. queenslandica* suggest that the latter scenario is correct. However, whether this is a general phenomenon among sponges (or other non-bilaterians) is still unknown.

Therefore, the aim of this study was to identify lincRNAs expressed during embryonic development in the calcisponge *Sycon ciliatum* (Calcaronea), a representative of one of the earliest evolved animal lineages. We have taken advantage of the existing large-scale RNA-Seq data [20] and systematically searched for long non-coding transcripts in different stages of embryogenesis. We identify 2421 transcribed lincRNAs and *in situ* hybridization (ISH) of selected representatives confirms that calcisponge lincRNAs are specifically and dynamically expressed in embryonic and somatic cells. More than 600 lincRNAs are specifically upregulated during embryogenesis. Finally, we have identified co-expressed modules of lincRNAs and coding genes that are active during specific stages of embryonic development and which are enriched for development-related functional categories. This study provides, to our knowledge, the first insight into the non-coding repertoire of calcisponges and supports the notion that RNA-based gene regulation was already present in the last common ancestor of all animals.

2. Methods

(a) Transcriptome assembly and identification of lincRNAs

Sycon ciliatum genome and protein-coding focused transcriptome assemblies have been previously described [20,21]. In this work, we have reassembled the transcriptome de novo from non-strand-specific poly(A)+ RNA-Seq reads using Trinity and detected protein-coding regions with Transdecoder with default parameters [22]. We chose de novo assembly over genome-driven assembly to alleviate effects of allelic variation between the genome and transcriptome (derived from different specimens) on on-genome alignment. Such variation influences on-genome alignment of short reads in much greater level than alignments of already assembled (and thus longer) transcripts. There were 46 967 unique transcripts identified as protein coding (minimum open reading frame (ORF) length 300 bp) and 46 824 as long non-coding (minimum length 600 bp; this stringent cut-off has been implemented to allow potential testing of expression by ISH in subsequent steps). The transcripts were aligned on the *S. ciliatum* genome assembly with Exonerate [23], which identified the structures of 26 349 coding and 21 680 non-coding genes. To ensure that the non-coding transcripts are truly not of coding origin (e.g. pseudogenes or remnants of retrotransposon activity), the 46 824 non-coding transcripts were used as queries in a BlastX search [24] against the NCBI RefSeq protein database (http://www.ncbi.nlm.nih.gov/refseq/). The Blast output was parsed with the BlastGrabber program [25], and sequences that gave a hit with an e-value of less than 10 were discarded. Such conservative e-value was chosen to ensure that no transcript of possibly coding origin was retained. The retained transcripts were translated in all six reading frames using transeq of the EMBOSS package [26] and used as queries in a HMMER search (e-value cut-off 0.01; [27]) against the PfamA database [28], as well as an additional Blastp search against the NCBI RefSeq database (e-value cut-off 10). The remaining transcripts were evaluated for protein-coding potential using the coding potential calculator [29]. All sequences with a coding potential score larger than 1 were discarded. In total, this left 10 548 transcripts from 6856 different genes that were putatively termed lncRNAs. As our assemblies are based on non-strand-specific libraries, differentiation between natural antisense transcripts and misassembled fragments of protein-coding genes is difficult. We have thus removed all sequences overlapping ORFs and introns of coding genes, leaving a dataset of 2421 intergenic lncRNAs (lincRNAs) for further analysis.

(b) *In situ* hybridization

To select candidate lincRNAs for ISH analysis, we used criteria which, in our hands, routinely give highly specific and robust expression patterns: expression level at least 40 counts in at least one library combined with at least 20-fold expression difference between any two stages. Of the 209 sequences satisfying these criteria, we have manually selected four transcripts representing diverse expression profiles (unique to early embryonic stages; peaking in the larvae; expressed throughout embryogenesis with or without expression in the larvae). Eight hundred to one thousand nucleotide fragments were amplified by PCR for each lincRNA and cloned using the pGEM-T easy vector system II (Promega, USA). Digoxigenin-UTP-labelled RNA probes were synthesized in both directions with SP6 and T7 RNA polymerases (Roche, USA) and cleaned using the RNeasy MinElute cleanup kit (Qiagen, USA). *Sycon ciliatum* specimens were collected in fjords near Bergen, Norway (N 60°27′33″, W 4°56′1″) between May and July 2013. The specimens were fixed, stored, hybridized and photographed as described in [30].

(c) Identification of independently regulated lincRNAs

To select lincRNAs with independently regulated expression, all lincRNAs with expression correlated with the nearest protein-coding gene neighbour (either upstream or downstream of the lincRNA) were discarded. Expression profiles of all identified coding and lincRNA genes across a range of developmental stages were calculated with use of the RSEM package [31] as described previously [20]. The neighbouring pairs were identified using closest-features in the BEDOPS toolkit v. 2.4.3 [32]. The Spearman correlation between pairs of a lincRNA and its neighbour gene was calculated in R v.3.1.2 [33], and p-values were corrected for multiple comparisons with the Benjamini–Hochberg (BH) procedure [34]. lincRNAs with a strong expression correlation with a protein-coding neighbour were discarded ($\rho \geq 0.6$, BH-adjusted p-value <0.05). Principal-components analysis (PCA) was performed with DESeq2 [35] on log-transformed normalized counts (using DESeq2 regularized log transformation).

Differential expression (DE) tests were performed using DESeq2 (Wald test with BH-adjusted p-values <0.1).

(d) Identification of co-expressed modules of lincRNAs and coding genes

To focus the analysis on relevant genes and to reduce the computational load, we only included the 10 560 coding genes which are differentially upregulated in any developmental stage compared with non-reproductive tissue, in addition to the 1853 identified lincRNAs. Furthermore, we discarded coding genes and lincRNAs with low variance between developmental stages; we required normalized counts higher than five in three or more samples, and we used only lincRNAs and coding genes with an expressional variance in the top 75% (variance calculated based on log-transformed ($\log_2(x+1)$ normalized counts). This filtering left 2615 transcripts (2421 coding genes and 194 lincRNAs). The module identification was done using the R package WGCNA v. 1.41 [36]. Modules were identified using the 'dynamic topological overlap matrix'-method and requiring a minimum module size of 30 (see the WGCNA manual). Briefly, Pearson correlations were calculated between all pairs and converted into an adjacency matrix using a power function (soft thresholding power 18). Adjacencies were converted into topological overlaps and clustered by hierarchical clustering in R. Modules were defined as branches cut-off using the dynamicTreeCut algorithm in WGCNA. Modules were assigned colour labels, which were then converted to letters from A-W (see the electronic supplementary material, figure S2).

(e) lncRNA blast search

The longest isoform from each of the 6856 lncRNA loci (repeats masked by Tandem Repeats Finder [37] and Repeat Masker [38]) was Blast searched (blastn word size 4, e-value cut-off 1×10^{-4}, minimum query overlap 25%) against the genomes of *Ciona intestinalis*, *Hydra magnipapillata*, *Nematostella vectensis*, *Amphimedon queenslandica*, *Oscarella carmela*, *Pleurobrachia bachei*, *Mnemiopsis leidyi*, *Trichoplax adhaerens*, *Salpingoeca rosetta*, *Sphaeroforma arctica* and *Capsaspora owczarzaki*, as well as the recently published *A. queenslandica* lncRNAs [15].

(f) Gene ontology analysis

All coding transcripts were searched for homologues against NCBI Refseq using BlastX. Blast results were imported into Blast2GO [39] and combined with conserved protein domain detection using InterProScan in Blast2GO to generate a gene ontology. In total, 10 552 genes were annotated. GO-enrichments of the different co-expressed modules were analysed by Ontologizer ([40]; topology-weighted method and p-value cut-off of 0.05). The GO-enrichment results were inspected manually and also visualized using the Enrichment Map Cytoscape plugin [41].

3. Results and discussion

(a) Thousands of lincRNAs are dynamically transcribed in *Sycon ciliatum*

An outline of the procedure aimed at identification of lincRNAs potentially involved in development of the calcareous sponge *S. ciliatum* is presented in figure 1. In the first step, we have used previously described non-strand-specific RNA-Seq datasets [20,21] to re-assemble the transcriptome, including non-coding sequences (our previous pipeline was focused on discovery of ORFs) and map it to the genome. A combination of Blast searches against reference databases, protein domain searches and ORF evaluation resulted in annotation of 2421 non-coding loci identified as putative lincRNAs.

Similar to what has been found in other studies, the lincRNAs were generally shorter than coding genes, with the majority of transcripts being below 1000 nts (electronic supplementary material, figure S1). In addition, the majority of lincRNAs were unspliced (i.e. single-exon transcripts), although a large number also contained multiple exons. To find out whether these genes might be developmentally regulated, both in terms of temporal and spatial expression, we have used a combination of bioinformatics and ISH. For the *in silico* part of the protocol, RNA-Seq libraries representing all key stages of oogenesis and embryonic development: vitellogenesis, fertilization, cleavage, cell differentiation and morphogenesis (referred to as preinversion and postinversion stages in calcaronean sponges) embedded in the maternal tissue, as well as free swimming larvae were used (figure 2*a*). We have visualized expression profiles of genes which, based on our experience in the *Sycon* model system, were likely to be robustly detected if studied by ISH (see Materials and methods). Among the 2421 queried putative lincRNAs genes, we selected four representatives with different developmental expression patterns for the subsequent ISH analysis. Consistent with the RNA-Seq data analysis, detection of all four probes revealed specific and unique expression patterns (figure 2*b–e*). In particular, the expression of scign021414 was limited to early stages of embryonic development and detected only in the embryonic cells until the preinversion stage, but not in surrounding maternal tissues (figure 2*b*). By contrast, scign009792 was not detectable in the oocytes or early embryos, but displayed strong expression in postinversion stage and larval micromeres (figure 2*c*). The remaining two genes were expressed in maternal cells only, or in both maternal and embryonic cells. scign011962 was detected in a variable fraction of choanocytes, especially those surrounding the oocytes and embryos, but not in the oocytes or embryos themselves (figure 2*d*). Finally, scign010682 was detected in a small number of unidentified small somatic cells, oocytes and early cleavage blastomeres (where it displayed nuclear and perinuclear localization), maternal cells ingressing into larval cavity during postinversion and in larval macromeres (figure 2*e*). Notably, in all cases, labelling was detected from one strand only, indicating unidirectional expression of all of the four lincRNAs studied by ISH. Thus, it appears that as in bilaterians, calcisponge lincRNAs display a striking variety of expression patterns, encompassing all embryonic cell types as well as multiple somatic cell types. In addition, their expression is clearly restricted to specific cell types and time points during development, which indicates that they are subjected to a tightly regulated transcriptional control.

(b) Hundreds of lincRNAs with independently regulated developmental expression

In bilaterian model systems, lncRNAs are often co-expressed with their coding genomic neighbours, which they sometimes overlap [42,43]. We have investigated genomic locations and expression of surrounding genes for the four selected examples of lincRNAs (figure 2*f–i*). As in the case of the expression patterns, the relationships between the position and expression of lincRNAs and their neighbours were varied. Three of the lincRNAs displayed no correlation of expression with their coding neighbours (figure 2*f,g,i*). Interestingly, the expression

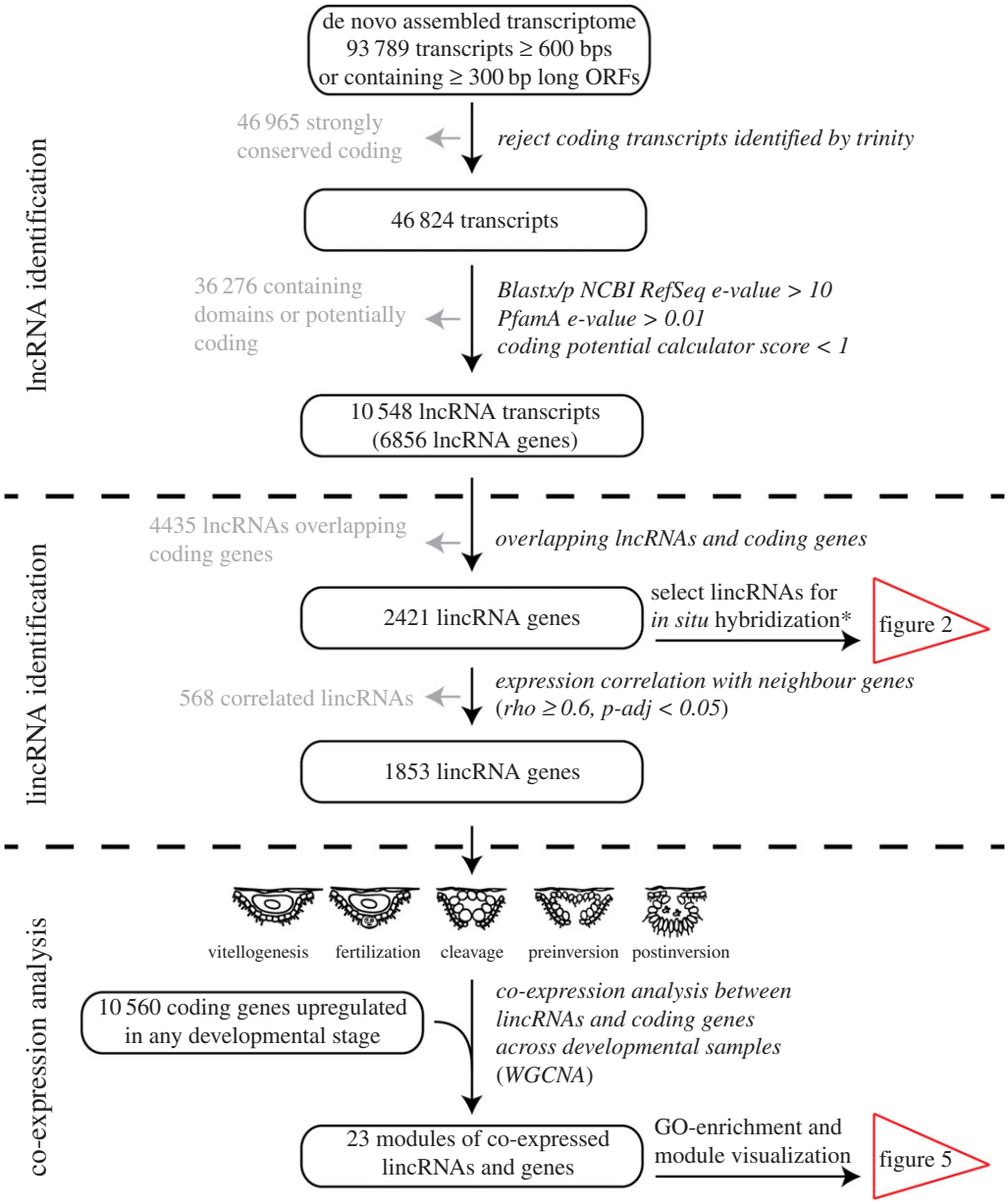

Figure 1. Overview of the filtering pipeline to detect lincRNAs in *Syon ciliatum*. The starting point of the analysis was a transcriptome assembled de novo from non-strand-specific pair-end RNA-Seq data (see the Methods section for details). Asterisk (*): criteria for selecting lincRNAs for *in situ* hybridization were expression level of at least 40 counts in at least one library combined with minimum 20-fold expression difference between any two developmental stages.

of scign009702 was moderately correlated ($\rho = 0.68$, $p < 0.001$) with scigt018127 transcribed in the opposite direction (figure 2*g*).

Given the diversity of the expression profiles and genomic organization of the large number of lincRNAs reflected by the four 'case studies' described above, we decided to systematically investigate lincRNA embryonic expression and co-expression with coding genes. To avoid artefacts caused by misassembly (such as misassembled fragments of UTRs) or erroneous transcription, and co-expression driven by genomic proximity rather than by functional relationships, we chose to focus this part of analysis on the lincRNAs with expression regulated independently of their neighbouring coding genes. We tested the 2421 lincRNAs for correlated expression with their closest protein-coding gene both upstream and downstream. Five hundred and sixty-eight lincRNAs were moderately or strongly correlated with a protein-coding neighbour ($\rho \geq 0.6$, BH-adjusted p-value < 0.05) and were discarded. Altogether, this left 1853 lincRNAs that we further analysed for potential association with development.

For this analysis, we have used the RNA-Seq libraries from the embryogenesis series for which biological replicates were available, as well as samples of sponges collected outside of the reproductive season and not containing any discernible oocytes or embryos [20]. Only mid-body slices of both reproductive and non-reproductive sponges were used as the oscular (apical) region of *S. ciliatum* has a different transcriptional makeup (as shown previously [20]).

We first wanted to know whether the expression of lincRNAs was structured according to the developmental stages and if so, whether this structure was similar or different to that of the coding genes. PCA demonstrated that expression of the lincRNAs was indeed strongly structured according to the different developmental stages, with non-reproductive tissue distant from all the other stages (figure 3*a*). This result is inline with our expectation that different pools of transcripts are active during the different stages of development. Notably, the structuring of lincRNAs expression seems to be very similar to that of the coding genes (figure 3*b*), and thus lincRNAs are likely to be involved in development similarly to coding genes.

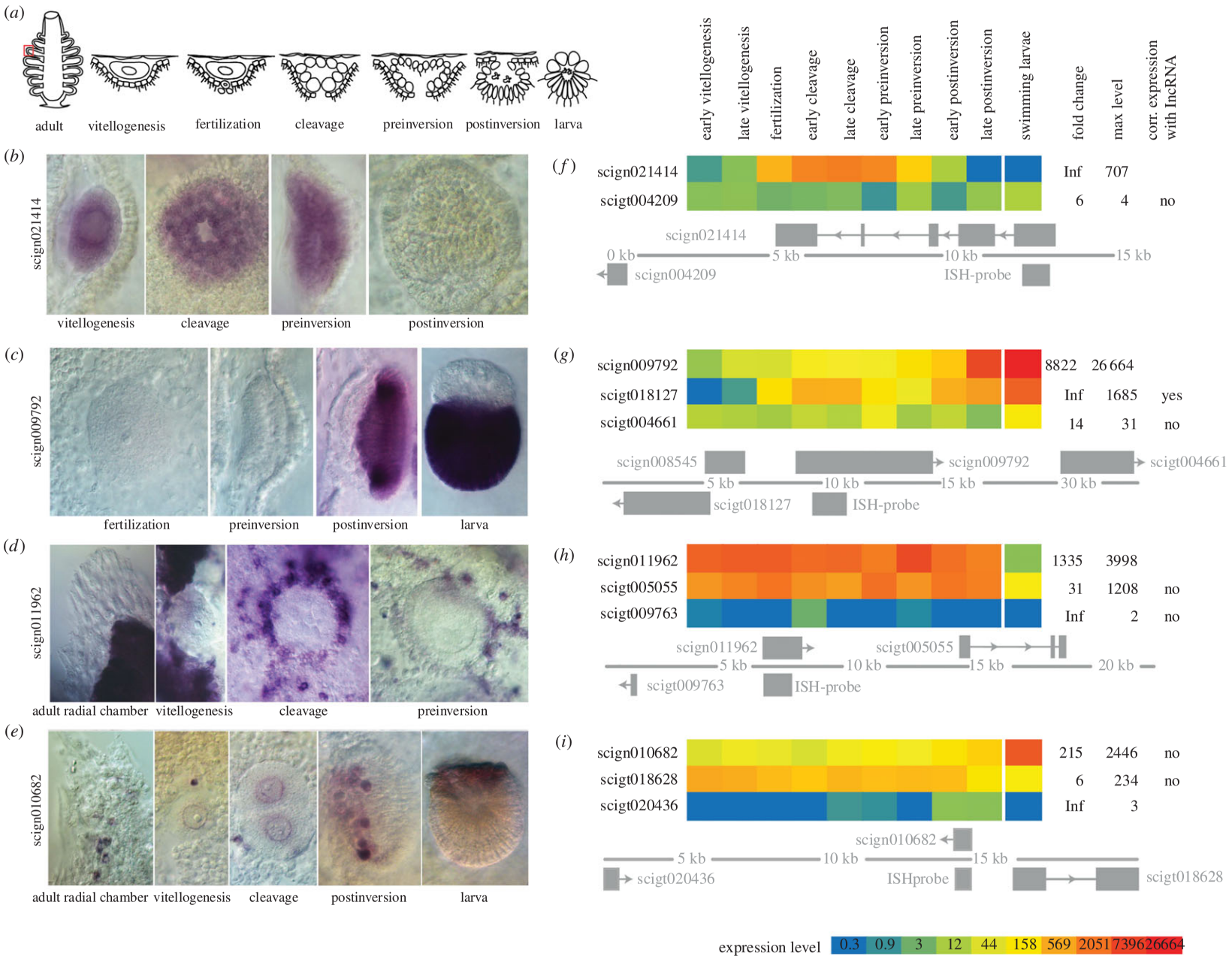

Figure 2. *In situ* hybridization (ISH) detection of developmentally expressed lincRNAs. (*a*) Overview of the different developmental stages used for the ISH and selection of lincRNAs for ISH. (*b–e*) ISH expression patterns of the lincRNAs during developmental stages. (*f–i*) Heatmap representation of expression of investigated lincRNAs and their nearest protein-coding neighbours on the genome, as well as a representation of the genomic localization. Significant correlations require $\rho \geq 0.6$ and p-value <0.05 (Spearman correlation).

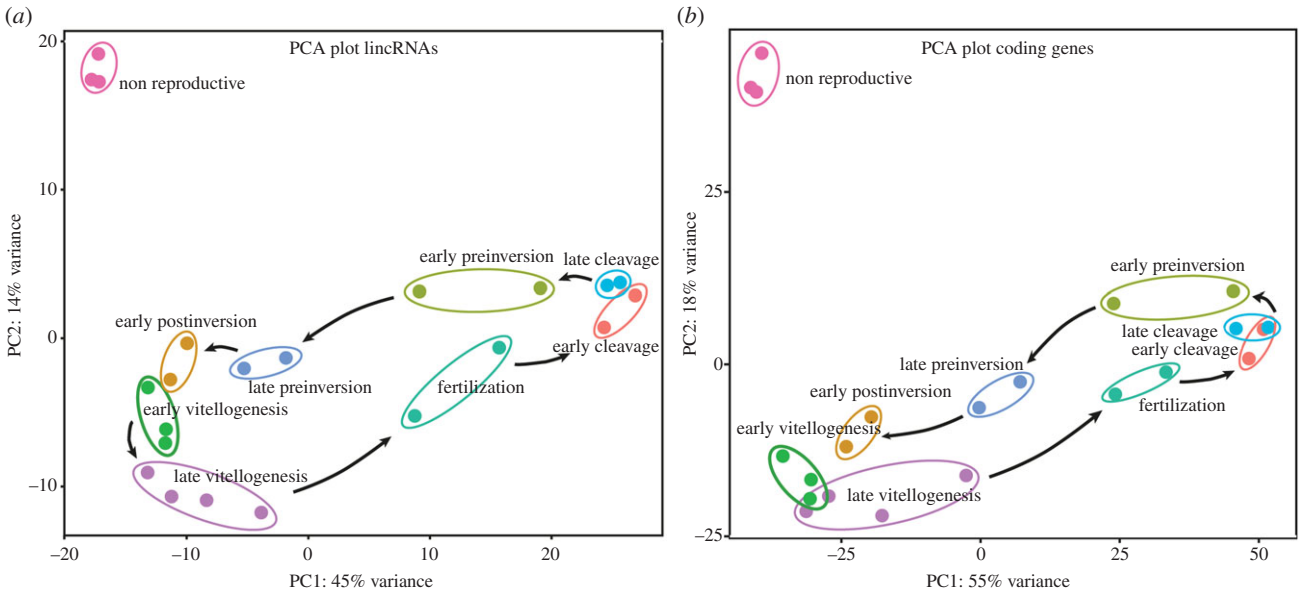

Figure 3. Principal-components analysis (PCA) plot. Plotting of the PCA on rlog-transformed gene expression counts from the different developmental stages of *Sycon ciliatum* of (*a*) lincRNAs and (*b*) coding genes. The analysis was done on the 500 most variable genes and the samples are plotted on their first two principal components. Each dot represents RNA-Seq data of a developmental sample.

To identify lincRNAs that were significantly upregulated during any of the developmental stages, we then tested for genes differentially expressed between non-reproductive samples and each developmental stage separately (figure 4). In total, 622 lincRNAs (33.6% of all independently regulated lincRNAs) were significantly upregulated in at least one of the developmental stages compared with non-reproductive tissue. In virtually all of the developmental stages, more than 200 lincRNAs were upregulated (except for 198 in early postinversion), with the cleavage stages displaying the highest numbers of upregulated lincRNAs (419 and 400). Successive stages of the development share the majority of upregulated lincRNAs, and 85 lincRNAs are upregulated across all developmental stages. On the other hand, only a small number of lincRNAs are uniquely upregulated in any developmental stage, with the highest number of unique lincRNAs (32) found during early cleavage. Thus, the cleavage stages appear to represent a period of very active transcription of a diverse pool of lincRNAs, perhaps in preparation to embryonic cell differentiation which will be occurring during subsequent developmental stages.

(c) lincRNAs are integral components of co-expressed gene modules including developmental regulatory genes

Gene regulatory networks and modules are central for the control and timing of organismal development. However, little is known whether non-coding genes are expressed in such modules. We therefore sought to identify modules of co-expressed coding genes and lincRNAs active during embryonic development. The co-expression analysis resulted in identification of 23 different modules (named A-W), with 21 of these including one or more lincRNA (figure 5 and the electronic supplementary material, figure S2). Two modules were almost uniformly expressed (indicated in figure 5 as the median of the normalized expression counts across all genes in a module) across development (J and K), and a few modules were restricted to a very narrow window during development, such as modules U and W only active during the latest stages of embryonic development (preinversion and postinversion).

The other modules showed two main patterns of expression; a large fraction of the modules seems to have the highest expression from fertilization to late cleavage or early preinversion stages (A–I). On the other hand, several modules (e.g. modules L–R) displayed a biphasic expression pattern almost opposite to modules A–I, with one peak during vitellogenesis and a second peak during morphogenesis stages (late preinversion and early preinversion). Given the fact that the second wave of oogenesis in *S. ciliatum* overlaps with these developmental stages, it is unclear whether this profile of expression is owing to expression in oocytes undergoing oogenesis only, or to expression present in oocytes, decreasing during cleavage and increased again in late preinversion stage embryos. Similar patterns of expression of protein-coding genes have indeed been previously observed, for example in the case of *SciBcatA* [20], although this gene is also strongly expressed in the somatic cells (choanocytes) of the adult tissue, and as such has not been recovered in our dataset of developmentally upregulated genes.

On the other hand, eight of the identified modules (but none of the strongly 'biphasic' modules) included protein-coding developmental genes (esp. components of the *Wnt* and *TgfBeta* pathways and transcription factors) with extensively studied expression patterns in *S. ciliatum* [20,21]. In addition, GO-term enrichment analysis (figure 5) indicated that several of the modules were particularly rich in terms related to developmental processes. For example, module I, including *SciFzdB*, *SciTGFBE* and *SciTbxB*, was particularly enriched in terms related to cell development and transcription factors. Module D, which included coding genes such as *SciNanos* and *SciNKC*, contained many genes related to cell differentiation and development, tissue and organ development and transcription regulation. Genes of both of these modules had a peak of expression during cleavage; with module D genes having a narrower peak of expression than module I. Similarly, a high fraction of genes included in module E (containing for example also *SciTGFBU*) have functions related to morphogenesis and organ development.

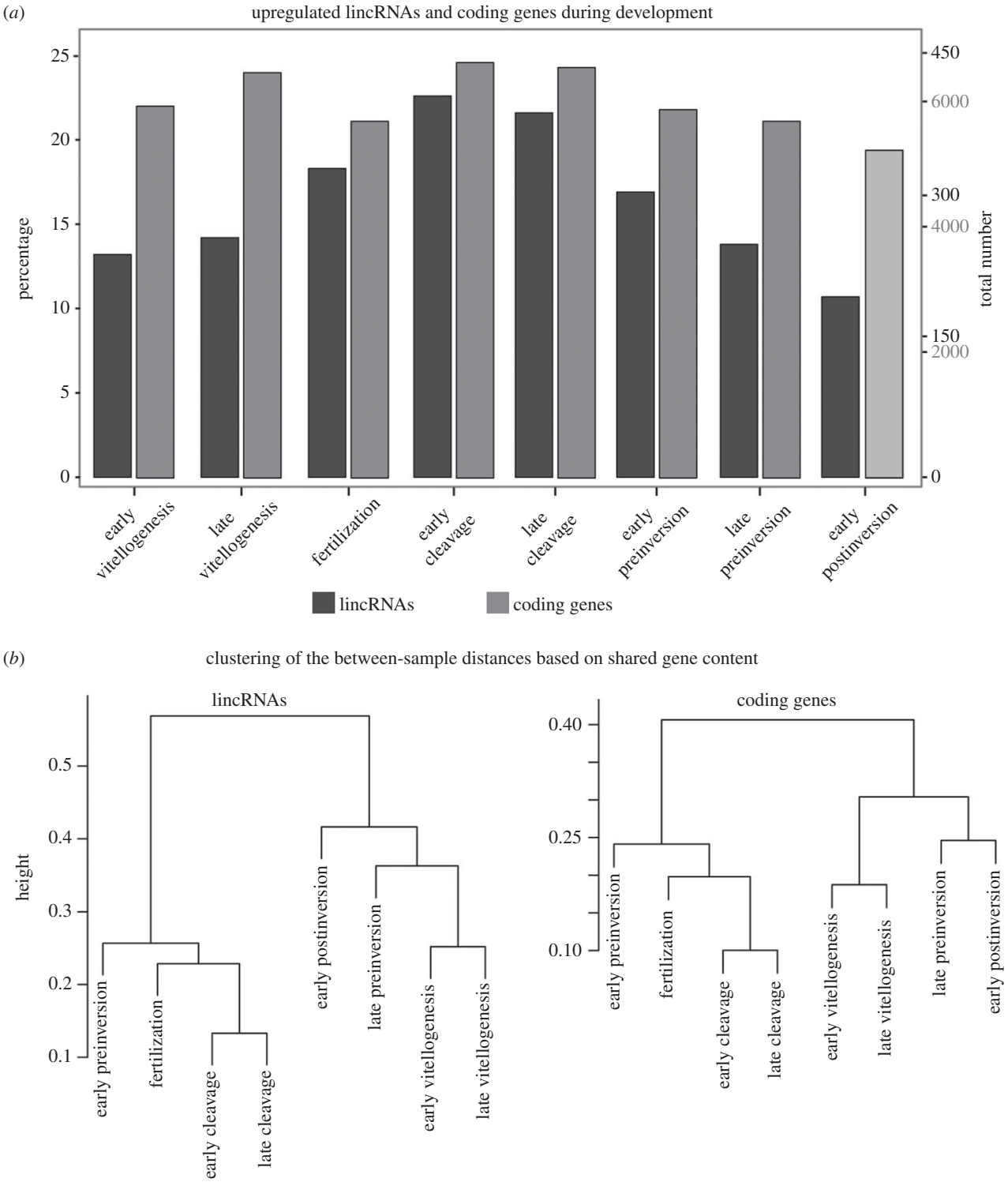

Figure 4. Differentially expressed lincRNAs and coding genes during development. (*a*) Histograms showing the number of significantly upregulated lincRNAs and coding genes between non-reproductive stages and reproductive stages. (*b*) Hierarchical clustering of the distances between developmental samples calculated on the basis of the shared number of upregulated lincRNAs or coding genes.

(d) lncRNAs as regulatory elements of animal development

It is becoming evident that lncRNAs are important for correct development of many animal lineages, for instance in mouse [18], zebrafish [11,16] and *C. elegans* [19]. Recently, lncRNAs expressed during development were also identified in the demosponge *A. queenslandica* [15]. In this study, we provide a first glimpse into the rich repertoire of regulatory RNAs involved in embryonic development of another early branching animal, the calcisponge *S. ciliatum*. This has important evolutionary implications; first of all, it suggests that using regulatory RNAs during early development is an ancestral feature of all sponges. Second, as both sponges and other animal lineages express lncRNAs during development, this feature was probably present already in the last common animal ancestor.

However, an important question is whether all sponges and other animals use homologues lncRNAs during development, or if they have acquired different types of lncRNAs during evolution. We did not identify any conserved lncRNAs between *S. ciliatum* and any other metazoan or non-metazoan opisthokont species. The lack of sequence similarity between

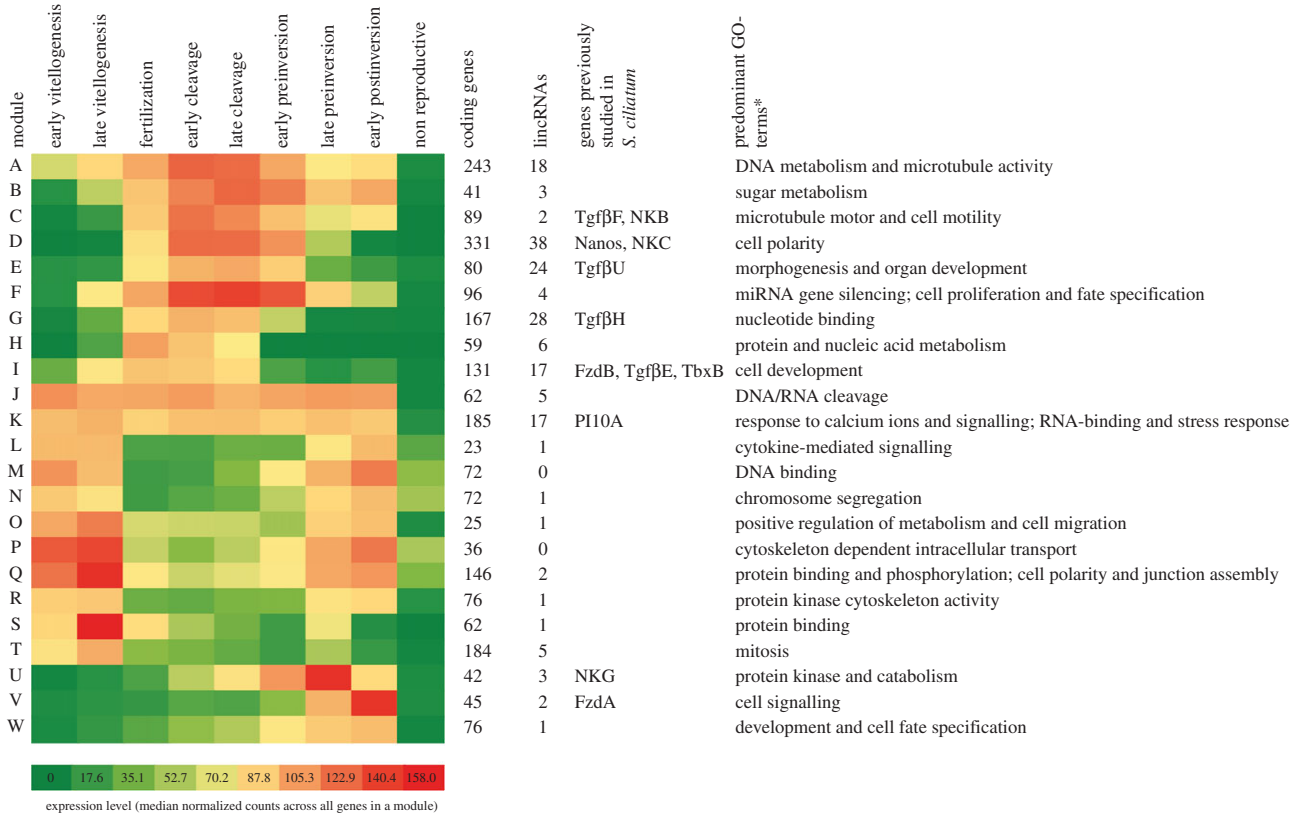

Figure 5. Overview of the co-expressed modules. The modules of co-expressed coding genes and lincRNAs are named from A to W. The heatmap is generated based on the median normalized expression values of all genes in a module. Asterisk (*): the predominant GO-terms are named on the basis of the major clusters of GO-terms in each module identified by the Enrichment Map analysis. In cases where no clusters were identified, the Ontologizer results were inspected manually.

lncRNAs across animal phyla (conserved lncRNAs have so far only been detected between vertebrate species [11,44]) suggests that these lncRNAs belong to different families, supporting the latter scenario. However, they could still have conserved secondary and tertiary structures, and thereby conserved function, despite being highly diverged on the primary sequence level.

The uncertain evolutionary history and the few functional studies undertaken so far makes it difficult to study lncRNA roles in an evolutionary developmental framework. One way to overcome this problem is to identify conserved modules, or networks, of co-expressed genes including lncRNAs. One such example could be the developmental lncRNAs co-expressed with *Frizzled B* (a key component of the Wnt-pathway) in both *A. queenslandica* [15] and *S. ciliatum* in this study (module I; figure 5).

Another challenge is that the availability of developmental transcriptome series is phylogenetically very patchy. Therefore, there is a need for high-quality staged transcriptome data from other deep-branching animal lineages, including ctenophores and placozoans. Such datasets might allow us to test whether, although lncRNAs are not conserved at the primary sequence level, they operate in deeply conserved gene regulatory networks.

Altogether, our work demonstrates that lncRNA expression during calisponge development is highly dynamic with restricted temporal and spatial patterns. Although it is uncertain whether these lncRNAs are homologous to those in other animals, the use of long non-coding RNAs in embryonic development is probably an ancestral feature of all animals.

Data accessibility. The following datasets are freely accessible on the Dryad Digital Repository (https://datadryad.org; doi:10.5061/dryad.v83fj): list of lincRNAs, PCR-primers used for ISH probe synthesis, coding and non-coding gene models, gene expression data, lists of co-expressed modules and the main R-commands used in this study.

Authors' contributions. J.B. conceived the study and participated in its design, computational analyses, molecular laboratory work and drafting of the manuscript. M. Adamski performed sequence assemblies and participated in the fieldwork, computational analyses and editing of the manuscript. R.S.N. participated in computational analyses. K.S.-T. participated in design of the study and editing of the manuscript. M. Adamska participated in design of the study, fieldwork, molecular laboratory work and analyses, drafting and editing of the manuscript. All authors read and approved the manuscript.

Competing interests. The authors declare that they have no competing interests.

Funding. J.B. is supported by the Norwegian Research Council, project no. 213707. R.S.N. and K.S.-T. are supported by the University of Oslo. M. Adamska and M. Adamski were supported by the Sars International Centre for Marine Molecular Biology, University of Bergen.

References

1. Fatica A, Bozzoni I. 2014 Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.* **15**, 7–21. (doi:10.1038/nrg3606)

2. Ingolia NT *et al.* 2014 Ribosome profiling reveals pervasive translation outside of annotated protein-

coding genes. *Cell Rep*. **8**, 1365–1379. (doi:10.1016/j.celrep.2014.07.045)

3. Rinn JL, Chang HY. 2012 Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem*. **81**, 145–166. (doi:10.1146/annurev-biochem-051410-092902)
4. Batista PJ, Chang HY. 2013 Long noncoding RNAs: cellular address codes in development and disease. *Cell* **152**, 1298–1307. (doi:10.1016/j.cell.2013.02.012)
5. Geisler S, Coller J. 2013 RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol*. 1–14. (doi:10.1038/nrm3679)
6. Koerner MV, Pauler FM, Huang R, Barlow DP. 2009 The function of non-coding RNAs in genomic imprinting. *Development* **136**, 1771–1783. (doi:10.1242/dev.030403)
7. Hacisuleyman E *et al.* 2014 Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat. Struct. Mol. Biol*. **21**, 198–206. (doi:10.1038/nsmb.2764)
8. Tsai M-C *et al.* 2010 Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689–693. (doi:10.1126/science.1192002)
9. Mercer TR, Mattick JS. 2013 Structure and function of long noncoding RNAs in epigenetic regulation. *Nat. Struct. Mol. Biol*. **20**, 300–307. (doi:10.1038/nsmb.2480)
10. Mattick JS, Rinn JL. 2015 Discovery and annotation of long noncoding RNAs. *Nat. Struct. Mol. Biol*. **22**, 5–7. (doi:10.1038/nsmb.2942)
11. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. 2011 Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**, 1537–1550. (doi:10.1016/j.cell.2011.11.055)
12. Ponjavic J, Oliver PL, Lunter G, Ponting CP. 2009 Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS Genet*. **5**, e1000617. (doi:10.1371/journal.pgen.1000617)
13. Cabili MN *et al.* 2011 Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*. **25**, 1915–1927. (doi:10.1101/gad.17446611)
14. Dunham I *et al.* 2012 An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74. (doi:10.1038/nature11247)
15. Gaiti F *et al.* 2015 Dynamic and widespread lncRNA expression in the sponge and the origin of animal complexity. *Mol. Biol. Evol.* **32**, 2367–2382. (doi:10.1093/molbev/msv117)
16. Pauli A *et al.* 2012 Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res*. **22**, 577–591. (doi:10.1101/gr.133009.111)
17. Sun J, Lin Y, Wu J. 2013 Long non-coding RNA expression profiling of mouse testis during postnatal development. *PLoS ONE* **8**, e75750. (doi:10.1371/journal.pone.0075750)
18. Sauvageau M *et al.* 2013 Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *eLife* **2**, e01749. (doi:10.7554/eLife.01749)
19. Nam J-W, Bartel DP. 2012 Long non-coding RNAs in *C. elegans*. *Genome Res*. **22**, 2529–2540. (doi:10.1101/gr.140475.112)
20. Leininger S *et al.* 2014 Developmental gene expression provides clues to relationships between sponge and eumetazoan body plans. *Nat. Commun*. **5**, 1–15. (doi:10.1038/ncomms4905)
21. Fortunato SAV *et al.* 2014 Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes. *Nature* **514**, 620–623. (doi:10.1038/nature13881)
22. Grabherr MG *et al.* 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol*. **29**, 644–652. (doi:10.1038/nbt.1883)
23. Slater GSC, Birney E. 2005 Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 1–11. (doi:10.1186/1471-2105-6-31)
24. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990 Basic local alignment search tool. *J. Mol. Biol*. **215**, 403–410. (doi:10.1006/jmbi.1990.9999)
25. Neumann RS, Kumar S, Haverkamp THA, Shalchian-Tabrizi K. 2014 BLASTGrabber: a bioinformatic tool for visualization, analysis and sequence selection of massive BLAST data. *BMC Bioinformatics* **15**, 128. (doi:10.1186/1471-2105-15-128)
26. Rice P, Longden I, Bleasby A. 2015 EMBOSS: the European molecular biology open software suite. *Trends Genet*. **16**, 276–277. (doi:10.1016/S0168-9525(00)02024-2)
27. Eddy SR. 2011 Accelerated profile HMM searches. *PLoS Comput. Biol*. **7**, e1002195. (doi:10.1371/journal.pcbi.1002195)
28. Punta M *et al.* 2012 The Pfam protein families database. *Nucleic Acids Res*. **40**, D290–D301. (doi:10.1093/nar/gkr1065)
29. Kong L *et al.* 2007 CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res*. **35**, W345–W349. (doi:10.1093/nar/gkm391)
30. Fortunato S *et al.* 2012 Genome-wide analysis of the sox family in the calcareous sponge *Sycon ciliatum*: multiple genes with unique expression patterns. *EvoDevo* **3**, 14. (doi:10.1186/2041-9139-3-14)
31. Li B, Dewey CN. 2011 RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323. (doi:10.1186/1471-2105-12-323)
32. Neph S *et al.* 2012 BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920. (doi:10.1093/bioinformatics/bts277)
33. R Core Team. 2014 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. See http://www.r-project.org/.
34. Benjamini Y, Hochberg Y. 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300.
35. Love MI, Huber W, Anders S. 2014 Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol*. **15**, 1–21. (doi:10.1101/002832)
36. Langfelder P, Horvath S. 2008 WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559. (doi:10.1186/1471-2105-9-559)
37. Benson G. 1999 Tandem Repeats Finder: a program to analyze DNA sequences. *Nucleic Acids Res*. **27**, 573–580. (doi:10.1093/nar/27.2.573)
38. Smit AFA, Hubley R, Green P. 2013 Repeat Masker Open-4.0. See http://www.repeatmasker.org/.
39. Conesa A, Götz S, Garcia-Gomez M, Terol J, Talon M, Robles M. 2005 Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676. (doi:10.1093/bioinformatics/bti610)
40. Bauer S, Grossmann S, Vingron M, Robinson P. 2008 Ontologizer 2.0 - a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics* **24**, 1650–1651. (doi:10.1093/bioinformatics/btn250)
41. Merico D, Isserlin R, Stueker O, Emili A, Bader G. 2010 Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE* **5**, e13984. (doi:10.1371/journal.pone.0013984)
42. Spurlock CF, Tossberg JT, Guo Y, Collier SP, Crooke PS, Aune TM. 2015 Expression and functions of long noncoding RNAs during human T helper cell differentiation. *Nat Commun*. **6**, 6932. (doi:10.1038/ncomms7932)
43. Tsoi LC *et al.* 2015 Analysis of long non-coding RNAs highlights tissue-specific expression patterns and epigenetic profiles in normal and psoriatic skin. *Genome Biol*. **16**, 1–15. (doi:10.1186/s13059-014-0570-4)
44. Necsulea A *et al.* 2014 The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635–640. (doi:10.1038/nature12943)